A New Computer-aided Technique for Qualitative Document Analysis

David Morris* & Nigel Ecclesfield

April 2010

**d.morris@coventry.ac.uk*

Introduction

Researchers are often faced with the problem of analysing large amounts of text with limited time resources. As Miles and Huberman note, the qualitative researcher is often presented with the challenge of converting "hundreds of pages of field notes to a final report" (1994:281). However, technology has overtaken even these challenges by encouraging the generation of greater and greater amounts of distributed textual data in a bewildering diversity of forms from the familiar journal articles, books and reports to the less familiar (for researchers at least) e-mails, blog postings, micro-blogging activity and so on. Researchers have always been faced with a choice between in-depth analysis of a small sample of the available relevant textual material and a more cursory analysis of larger amounts of textual data. The development of computer-aided qualitative data analysis software[1] (CAQDAS) has shifted the balance by providing the means to classify, organise and analyse data very effectively and, perhaps more importantly, to search across data quickly, accurately and comprehensively to identify themes.

However, the existing techniques are still time-consuming and may not be viable in where there is either a large amount of data to be analysed and/or speed is required. This is particularly the case where it is felt necessary or desirable for multiple researchers to analyse text independently in order to help eliminate researcher bias and improve reliability. In many cases, for example in PhD work, this may not prove to be a viable proposition. To date the use of computers has been somewhat limited to searching, counting, coding and retrieval activities. However, there have been some recent developments in the field of concept mapping. By coding we mean the activity of sorting textual data into categories which have been predefined by the researcher according to a set of characteristics (usually phrases or words) attached to each category. More recently, software (see, for example, Wordle[2]) has appeared which presents visual summaries of counts of words which are not based on predefined categories. Such "word clouds" can be very effective in giving a visual snapshot of large amount of text. Categorising (and coding as the means of achieving it) re-orders text (data) into a similarity-based ordering which replaces the original ordering. Computer programmes have proved very effective aides in this re-ordering process.

The use of computers in qualitative data analysis is not new (Kelle, 2004), although many would argue that it has inbuilt limitations:

Computers.....maybe inappropriate for the analysis of texts, can still be helpful for the organisation and analysis of textual data (Kelle, 2004:446).

Bryman (2004: Ch. 18) defines qualitative content analysis as the searching out of underlying themes in the documents being analysed.

However, the most important research process is one of connecting data. Maxwell and Miller (2009:467) note that:

.....connecting analytic strategies do not simply preserve data in their original form. Instead, they are ways to analyse and reduce data. This is generally done by identifying key relationships that tie

[1] For example Atlas/ti, NUD*IST and NVIVO.
[2] http://www.wordle.net

the data together into a narrative or sequence and eliminating information that is not germane to these relationships....

As Maxwell and Miller suggest, computer programmes have not been very useful in the process of connecting data except through their ability to help organise, track, search and relate data elements to each other.

An alternative approach

The approach we describe here is not intended to supplant the use of existing CAQDAS techniques and methods. However, what we propose does add to the existing range of techniques available and may merit a place in the qualitative researcher's toolkit. In terms of the trade-off noted above, the technique lies at the end of the spectrum where very large amounts of data can be analysed very quickly but at less depth. The technique also concentrates on documents as content and sources of evidence rather than as topics of research in their own right using methods such as discourse analysis (see Prior, 2008). A further key feature is that categories emerge from iterative analysis of the document itself rather than being defined by the researcher and then "imposed" on the document. This is in stark contrast to conventional approaches to content and textual analysis where it is argued that categorisation is the "most important part of content analysis because it reflects the purposes of the research and the theories underlying it"; in consequence categories need to be "objectively defined" (Anderson, 1997:341).

The core of the software used is Cirilab's[3] "Speed Read" and "Knowledge Map". Speed Read is based on three sequential activities. The initial stage is knowledge indexing which captures, organises and classifies unstructured text to provide a means of easily navigating through it. This stage is accomplished by the software itself; the researcher does not have to pre-specify any codes, categories or key words or phrases to be used as the basis of classification. The software identifies categories based on frequency of occurrence. Secondly thematic indexes are created; these identify the most important themes within the text[4]. Finally a complex set of algorithms[5] is applied to identify relationships between the themes through identifying recurring patterns within the text. This element of the software is a Cirilab proprietary system; however a major criterion for grouping themes together is proximity within the text. This pattern of themes is known as a document's "knowledge signature". The signature also includes a summary of the document. Users can use the knowledge signature to all the elements of a document which contribute to a given theme. This allows discussion of themes to be easily illustrated using quotes from the original document. Knowledge Maps combine individual knowledge signatures to give an overall picture of a library of documents. The Document Navigator can then be used to undertake thematic searches of document libraries and drill down through the themes.

Knowledge Manager performs the same types of functions but can deal with a library of documents rather than a single one in the case of Speed Read. The first stage creates a Knowledge Signature for each document in the library using the same methodology as Speed Read. Knowledge Signatures for

[3] http://www.cirilab.com

[4] This is a convenient simplification. The software uses sophisticated "stemming" algorithms in its identification of categories.

[5] Known as "Multidimensional Semantic Spatial Indexing". One of the principal algorithms used is Genex (see Turney, 2000).

each document can be viewed independently. However the system can also combine the document-level Knowledge Signatures into a Knowledge Map of the entire collection. The Document Navigator is a thematic map of the entire collection and allows themes to be identified and displayed across the document collection. The navigator also allows easy retrieval of the parts of the documents where the themes occur for comparison and analysis. Again this makes the identification of relevant quotes very easy.

These Cirilab sytems can be integrated with other systems to provide a framework for electronic knowledge management (Roman and Spearing, 2008).

The main current research use to date has been to help reduce the information overload when analysing data which might help in dealing with emergencies (Roman et al., 2008). A useful introductory review of the software has been provided by Farris (2007).

Advantages and disadvantages

This approach has a number of potential advantages and uses:

- Where large amounts of text is involved
- Where speed is necessary
- It is relatively low cost, both in terms of software license fees and in time spent learning to use the system[6]
- As a comparison and complement to other approaches to analysing the same data (for example manual coding and interpretation)
- As part of a triangulation research design
- As an approach to eliminating researcher bias[7] principally by removing the responsibility for coding from the researcher

The major disadvantages are:

- The method, as yet, has not been compared with or benchmarked against other more familiar approaches
- The computer software has difficulty in dealing with synonyms which are not variants of the same word (writers may use synonyms to make their styles more interesting for the reader, even if there is no other gain to doing so)
- It places an intermediary (or, as some would argue, a barrier) between "researchers and their data"[8]
- It diverts researcher attention away from the processes of coding and categorisation which are potentially at the core of some research
- Use is limited to text in the narrow sense

[6] The software used is essentially aimed at the consumer market although it is part of a much wider knowledge management suite aimed at corporate use.

[7] The limitations of computer applications (the inability to think creatively about the data) can also be a strength (the computer does not approach the data with any preconceptions). The software "let's the document speak for itself".

[8] Note the use of "their". Many writers argue that qualitative document analysis (QDA) relies on researchers being immersed in their data and subject matter. See, for example, Altheide at al. (2008).

The technique is most likely to be of use within an exploratory research design where multiple methods are being employed (Plano Clark et al., 2008). On the other hand, the technique is potentially quite limited when it comes to theory-building rather than data exploration.

Our experiences of using the software

We have used this approach in a number of ways:

- Analysis of project bids to check concordance with bid requirements
- Analysis of project bids to identify common themes
- Analysis of strategic plans
- To aid PhD students in preparing the final versions of dissertations

To date our only "test" of the approach has been face validity. However, in most instances, this has been high. Through a process of trial and error we have found that the method works best when:

- There are large amounts of text to be processed
- All structural cues (tables of contents, headings, sub-headings, headers, footers etc) are removed
- All diagrams, lists of references etc. are removed

Whilst these processes might seem to be consuming, converting the text to RTF format before analysis will accomplish much the same ends. These practices are consistent with the notion that the software does not attempt to interpret the text but uses a natural language processing approach. Categorisation is done after the fact.

Given that we are not aware of any systematic comparisons of the outputs from this approach with those derived from analysis of the same sources by other tried and tested means (benchmarking), there is a need to exercise caution in interpreting the results. However, the potential advantages of the technique (including resource leanness and speed) could render it a valuable component of a multi-method research design.

Representing the data

The mind (or "concept") map has shown itself to be a very useful means of identifying key relationships between concepts, ideas, actions and dependencies in a wide variety of situations including planning research activities, project development and as a writing aid. Mind maps also reduce data very effectively into an often compelling visual representation of the interconnections. However, they have been mostly used in ex ante (planning) situations, that is as inputs to research or other projects rather than as outputs from data analysis. On the other hand there seems little reason to suppose that they would not be useful as a means of summarizing large amounts of data and expressing the key relationships (connections) between different elements (categories) within that data. This would be an ex post rather than ex ante use. Mind maps take us several steps beyond the visual representations of tag or word clouds.

An illustrative example

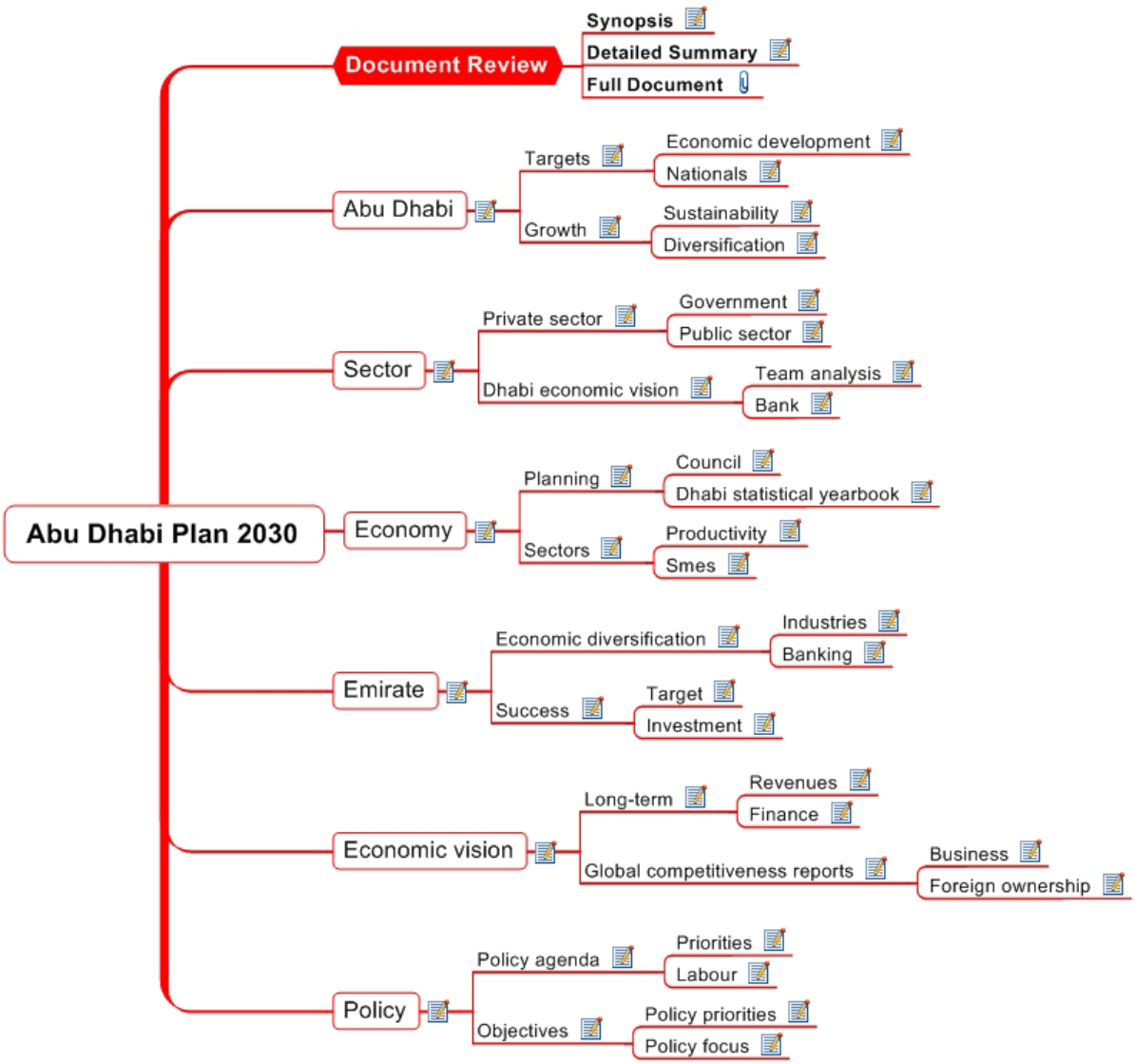
As a part of a project on the approaches being adopted by a number of countries towards strategic planning we wished to analyse the Abu Dhabi plan[9], a document of approximately 31,000 words. The resulting themes (plus all the underlying text summaries, indexes etc) are then exported to a mind mapping software system[10] and a mind map is created from the relationships identified. The entire operation took only a few minutes from downloading the document to generating the mind map. An image of the resulting output is presented below.

The mind-map reflects the default 6-2-2 structure which the Cirilab system uses. Thus there are 6 principal themes (the top branch of the mind-map contains the document review). Each principal theme is divided into two sub-themes which are, in turn, divided into two more. This gives a total of 42 themes. The 6-2-2 structure has been found to provide a useful compromise between usefulness and comprehensiveness (in most cases themes beyond the 42 mark tend to be based on very few references).

Mind maps provide a very useful means of analysing documents. The graphical representation allows the researcher to identify areas of similarity and difference much more easily than direct comparison of the underlying text would allow. The mind map also allows a degree of interactivity which promotes active exploration of the underlying themes. In the original (full) version of the mind map it is possible to click on each of the "page and pencil" icons for a complete summary of all the references to the original text which have been used to generate the particular topic. Arguably this promotes deeper exploration of the underlying document than would otherwise have been possible and potentially invites researcher immersion in the data.

[9] http://www.usuaebusiness.org/view/images/uploaded/Abu%20Dhabi%202030%20Vision%20Report.pdf

[10] ThemeReader from MindSytems; http://www.mindsystems.com.au/products/themereader


Abu Dhabi Plan 2030
Document Review
Synopsis
Detailed Summary
Full Document
Abu Dhabi
Targets
Economic development
Nationals
Growth
Sustainability
Diversification
Sector
Private sector
Government
Public sector
Dhabi economic vision
Team analysis
Bank
Economy
Planning
Council
Dhabi statistical yearbook
Sectors
Productivity
Smes
Emirate
Economic diversification
Industries
Banking
Success
Target
Investment
Economic vision
Long-term
Revenues
Finance
Global competitiveness reports
Business
Foreign ownership
Policy
Policy agenda
Priorities
Labour
Objectives
Policy priorities
Policy focus

References

Altheide, David; Coyle, Michael; DeVriese, Katie and Schneider, Christopher (2008). Emergent Qualitative Document Analysis IN Sharlene Nagy Hesse-Biber and Patricia Leavy (eds.) **Handbook of Emergent Methods.** New York: The Guilford Press.

Anderson, J (1997). Content and Text Analysis IN John P. Keeves (Ed) **Educational Research, Methodology and Measurement: An International Handbook** 2nd. Edn. Oxford: Pergamon

Bryman, Alan (2004). **Social Research Methods**. 2 edn. Oxford: Oxford University Press

Farris, Dale (2007), *Cirilab Software*. Available at: http://www.gtpcc.org/gtpcc/cirilab.htm

Kelle, Udo (2004). Computer-assisted qualitative data analysis IN Clive Seale, Giampietro Gobo, Jaber F. Gubrium and David Silverman (eds) **Qualitative Research Practice**, London: Sage Publications

Maxwell, Joseph A and Miller, Barbara A (2008). *Categorizing and Connecting Strategies in Qualitative Data Analysis*, IN Sharlene Nagy Hesse-Biber and Patricia Leavy (eds.) **Handbook of Emergent Methods.** New York: The Guilford Press.

Miles, M.B. and Huberman, A.M. (1994). **Qualitative data analysis: An expanded source book** (2nd. Ed.) Thousand Oaks, CA: Sage Publications

Plano Clark, Vicki L; Creswell, John W; Green, Denise O'Neil and Shope, Ronald J (2008). *Mixing Quantitative and Qualitative Approaches: An Introduction to Emergent Mixed Methods* Research IN Sharlene Nagy Hesse-Biber and Patricia Leavy (eds.) **Handbook of Emergent Methods.** New York: The Guilford Press.

Prior, Lindsay (2008). Researching Documents: Emergent Methods IN Sharlene Nagy Hesse-Biber and Patricia Leavy (eds.) **Handbook of Emergent Methods.** New York: The Guilford Press.

Roman, Jorge H; Collins, Linn Marks; Mane, Ketan K; Martinez, Mark L.B.; Dunford, Carolyn E and Powell, James E. (2008). Reducing *Information Overload in Emergencies by Detecting Themes in Web Content*, IN F. Fiedrich and B. Van de Walle (eds) **Proceedings of the 5th. International ISCRAM Conference**, Washington, DC: May 2008.

Roman, Jorge H and Spearing, Shelly A (2008). **electronic Knowledge Management (eKM) Today**, DigIn Digital Preservation Conference, June 4-6.

Turney, P.D. (2000). Learning Algorithms for Keyphrase Extraction, **Information Retrieval** 2 (4): 303-336.